CS 70 Spring 2004 — Discrete Mathematics for CS — Papadimitriou/Vazirani — HW 13

Due on Monday, May 3 at 5pm (283 Soda)

This week's homework assignment consists of two programming exercises. You are free to write your programs in any "standard" programming language (by "standard," we mean any language for which there is a compiler or interpreter on the instructional machines). Your submissions should consist of all output files, plots, etc. that you produce when working on these problems, along with a printed copy of your source code.

1. **(50 pts.) Distributions in the limit**
   For any (nontrivial) probability space Ω of your choice, define some random variable X on Ω (do not choose a random variable that is too simple—e.g., a constant r.v.–but you don't have to choose one that's very complicated either). For this problem, you will verify the idea that "all distributions approach the normal distribution asymptotically."

   We can define the random variable A_n on Ω by

   $$A_n = \frac{1}{n}\sum_{i=1}^{n} X;$$

   intuitively, A_n is the average of X. Write a program that samples from the distribution of A_n as follows:

   Write a program that on input $n \geq 1$ and $k \geq 1$ does the following: samples n times from Ω, averages the values of X on those sample points, and repeats this k times. This is equivalent to sampling the random variable A_n k times. Plot the distribution for various values of k—in particular, choose k large enough so that the distribution is clear.

   The central limit theorem says that the distribution of A_n should approach the normal distribution as $n \to \infty$. How large does n have to be before you see noticeable "convergence"?

2. **(50 pts.) The Polya Process**

   (a) Recall that in the Polya Urn Process, the probability that the next ball goes into bin k is proportional to the number of balls in k. Analyze the Polya Urn process with m balls in n urns (initially each urn has one ball). Show that the number of balls in the urns can be described by randomly interspersing $n-1$ red balls among the m balls arranged in a line. The balls between the $k-1$st and kth red ball are the balls in the kth urn.

   (b) Now let $m = cn$ for various constants c (e.g., $c = 3$). Write a program that simulates the process and let $f(k)$ be the number of balls in the kth largest urn. What is the function f? [Hint: plot $f(k)$ versus k on a log-log plot. Why would a straight line indicate a Zipf like distribution?]